

*EFFICIENTLY ESTABLISHING CONCEPTS OF INFERENTIAL
STATISTICS AND HYPOTHESIS DECISION MAKING THROUGH
CONTEXTUALLY CONTROLLED EQUIVALENCE CLASSES*

DANIEL M. FIENUP AND THOMAS S. CRITCHFIELD

ILLINOIS STATE UNIVERSITY

Computerized lessons that reflect stimulus equivalence principles were used to teach college students concepts related to inferential statistics and hypothesis decision making. Lesson 1 taught participants concepts related to inferential statistics, and Lesson 2 taught them to base hypothesis decisions on a scientific hypothesis and the direction of an effect. Lesson 3 taught the conditional influence of inferential statistics over decisions regarding the scientific and null hypotheses. Participants entered the study with low scores on the targeted skills and left the study demonstrating a high level of accuracy on these skills, which involved mastering more relations than were taught formally. This study illustrates the efficiency of equivalence-based instruction in establishing academic skills in sophisticated learners.

Key words: college students, contextual control, inferential statistics, instruction, stimulus equivalence

The present study was designed as a preliminary step toward developing instructional procedures based on stimulus equivalence to teach college students the beginning concepts of inferential statistics and hypothesis decision making. The study was prompted by our experience with students in sophomore-level research methods courses routinely scoring lower in the inferential statistics unit than in other course units. This experience is consistent with the widely held belief that inferential statistical concepts are among the most difficult to teach in an undergraduate psychology curriculum (e.g., Knowles, 1974; Kranzler, 2007).

Interventions using stimulus equivalence to teach high-level skills to typically functioning individuals may play an important role in the translational agenda of behavior analysis (e.g.,

Perone, 2002). Although laboratory studies have established complex equivalence-based repertoires in typically functioning adults using arbitrary stimuli (e.g., Belanich & Fields, 2003; Dougher, Perkins, Greenway, Koons, & Chiasson, 2002; Griffiee & Dougher, 2002; Lane, Clow, Innis, & Critchfield, 1998; Pilgrim & Galizio, 1995), there have been few attempts to employ instruction based on stimulus equivalence and other stimulus relations (hereafter called equivalence-based instruction; EBI) to enhance the teaching of high-level learners. Most stimulus equivalence research that bridges basic science and application has been conducted with individuals with intellectual challenges (e.g., de Rose, de Souza, & Hanna, 1996; Lane & Critchfield, 1998; Sidman & Cresson, 1973; Taylor & O'Reilly, 2000) or with typically developing children (Connell & Witt, 2004; Lynch & Cuvo, 1995).

There exist few published reports in which EBI was used to teach sophisticated academic concepts to advanced learners. Ninness et al. (2005) used a match-to-sample procedure to teach formula-graph functions to college students who lacked relevant skills (see also related reports by Ninness et al., 2006, 2009). During A→B training, students learned to match

This research was conducted at Illinois State University as part of the first author's requirements for a doctoral degree in school psychology. We thank Larry Alferink, Dawn McBride, Gary Cates, and James Dougan for comments and suggestions, and Daniel Covey for assistance in conducting the research study.

Address correspondence to Daniel M. Fienup, who is now at the Department of Psychology, Queens College, 65-30 Kissena Blvd., Flushing, New York 11375 (e-mail: daniel.fienup@qc.cuny.edu).

doi: 10.1901/jaba.2010.43-437

factored (reduced) formulas (B) to standard, nonreduced formulas (A). During $B \rightarrow C$ training they learned to match graphed functions (C) to factored formulas (B). Without additional training, all 11 participants showed the emergence of untaught relations ($A \rightarrow C$ and $C \rightarrow A$), in which they were able to match graphs to standard equations and vice versa. These effects generalized to new types of mathematical functions not used in training. In a similar vein, Fields et al. (2009) examined EBI of concepts related to statistical interactions. Participants were taught to conditionally relate graphs to behavioral examples, behavioral examples to interaction terms (e.g., crossover, interaction), and interaction terms to definitions. From this instruction, participants were able to match definitions to graphs and generalize this responding to novel graphs and novel test formats.

These interventions illustrate an important form of instructional efficiency with relatively sophisticated students in that they reliably instated more relations than were explicitly taught. These cases fuel optimism that EBI can enhance the learning experiences of advanced learners in the same way they appear to do for learners focused on elementary academic skills.

As far as we are aware, previous applied stimulus equivalence studies, whether focused on advanced or rudimentary instruction, have all involved unconditional equivalences. For example, in the fraction–decimal equivalences taught by Lynch and Cuvo (1995), 0.5 and 1/2 always “go together” in the same way. However, the logic behind hypothesis decision making is conditional, in that decisions about hypotheses require attention not only to the direction of an experimental effect but also to statistical significance in many cases. Knowing whether an effect is in the direction predicted by the scientific hypothesis is necessary, but not sufficient, to support hypothesis decisions; also required is attention to whether the results are statistically reliable (Huck, 2000). To cite an incomplete illustration of this conditional

reasoning, given an effect that matches the direction predicted by a one-tailed scientific hypothesis, the effect “goes with” *reject null hypothesis* only if the effect is statistically significant; otherwise, the effect “goes with” *fail to reject the null hypothesis*.

In the stimulus equivalence literature, conditional reasoning is referred to as *contextual control* (Sidman, 1994), and laboratory studies show that it can be established via procedures of stimulus equivalence (e.g., Dougher et al., 2002; Griffie & Dougher, 2002; Rehfeldt, 2003; Steele & Hayes, 1991). An example described by Bush, Sidman, and de Rose (1989) illustrates how contextual control operates. In the context of *nationality*, Claude Monet belongs among people of French origin, including nonpainters Charles DeGaulle and Pope Gregory XI. In the context of *profession*, Monet belongs among painters who include non-French artists such as Grandma Moses and Michaelangelo. Here, the contextual cues may be said to control or “switch on” and “switch off” Monet’s membership in various stimulus classes (i.e., nonartists from France and artists who are not French).

Although no previous instructional application of stimulus equivalence technology has taught concepts that required conditional reasoning, the principles that apply to abstract stimuli should apply to contextually controlled learning about stimuli of everyday relevance (e.g., Keenan, McGlinchey, Fairhurst, & Dillenburger, 2000; Kohlenberg, Hayes, & Hayes, 1991; Mattaini, 1999), an assumption that our recent pilot work supports (Fienup, Critchfield, & Covey, 2009). The purpose of the present study was to use contextual control to teach students the conditional application of concepts of statistical significance and hypothesis decision making.

METHOD

Participants and Setting

Thirteen undergraduate college students participated for up to 2 hr after providing informed

consent. In exchange for participating, they received vouchers that could be exchanged for bonus credit in psychology courses. Volunteers were retained in the study if they scored below 70% on the pretest of both Lesson 1 and Lesson 2. This criterion was used in an attempt to avoid ceiling effects that might obscure any evidence that the experimental procedures promoted learning. It is important to note, however, that each test encompassed several types of relations, and on a given pretest a student could score $\geq 70\%$ correct on one type of relation and still score under 70% for the test overall. In such cases, the volunteer was retained in the study, because the goal of recruitment was to identify individuals who could profit from the lessons rather than simply those who knew nothing whatsoever about statistical inference.

Three volunteers were dropped from the study after scoring too high on pretests, and no data are reported for these individuals. The remaining 10 student participants (six women, four men) ranged in age from 18 to 28 years ($M = 20.1$, $SD = 2.96$) and reported college grade-point averages ranging from 1.40 to 3.70 ($M = 2.94$, $SD = 0.72$). Only one student reported any prior experience in a statistics or mathematics course relevant to the topic of instruction. At the time the study was conducted, Student 10 was enrolled in an introductory statistics course but was retained after scoring poorly on pretests.

Students completed all study tasks during a single visit to a classroom that was equipped with 30 computers arranged in four rows of seven or eight workstations each. Each student worked on an IBM-compatible desktop computer (with a 15-in. flat panel monitor, keyboard, and mouse) that ran on the Microsoft Windows XP operating system and controlled study events automatically via a custom-written program that was created with Visual Basic 2005 (Dixon & MacLin, 2003).

General Procedure

Instructions. The experimenter asked participants to complete three computerized lessons

on inferential statistics and hypothesis decision making that were described as in development for future use in a university course. The experimenter told participants that each lesson included a pretest, learning phase, and posttest; that a score of at least 90% was needed on each posttest to progress through the study; and that dismissal from the study would occur after 2 hr or completion of all three lessons, whichever came first. Because instructions did not describe the learning stimuli or relations among them, they are not reproduced here, but are available from the first author.

Overview of lessons. For each of the three lessons, students completed a pretest, training, and a posttest, each of which is described in detail below. The first lesson was based on concepts of statistical significance and taught relations between stimuli that in the present tables and appendixes are labeled A, B, and C for convenience (students were not exposed to this notation). The second lesson was based on concepts of hypothesis decision making and taught relations between stimuli that in the tables and appendixes are labeled D, E, and F. The third lesson taught contextual relations involving the previously mastered stimuli. In all lessons, the pretest and posttest were identical and incorporated all of the potential relations among the lesson's stimuli. In all lessons, the training phase incorporated two or more blocks of trials or learning units that had to be mastered separately. See Table 1 for an overview of the stimuli and notation used in this study.

Learning stimuli and match-to-sample procedure. The learning stimuli were presented on the computer screen in black font in white boxes (7.6 cm by 7.6 cm) that were arranged with the sample stimulus at the top of the screen with three comparison stimuli below (see Figure 1 of Fienup et al., 2009). Each stimulus was displayed in Times New Roman font in a size that largely filled the white box (range, 20- to 40-point font), although during Lesson 3 the

Table 1
Stimuli Used in the Study and Notation

| Notation | Set 1 | Set 2 |
|----------|---|--|
| A | Low p value | High p value |
| B | Statistically significant | Not statistically significant |
| C | $p \leq .05$ | $p > .05$ |
| ↑ D | Scientific hypothesis: the IV will increase the DV Results: the DV increased | Scientific hypothesis: the IV will increase the DV Results: the DV did not increase |
| ↓ D | Scientific hypothesis: the IV will decrease the DV Results: the DV decreased | Scientific hypothesis: the IV will decrease the DV Results: the DV did not decrease |
| ↕ D | Scientific hypothesis: the IV will change the DV Results: the DV changed | Scientific hypothesis: the IV will change the DV Results: the DV did not change |
| E | Consistent with scientific hypothesis | Not consistent with scientific hypothesis |
| F | Reject null hypothesis | Fail to reject null hypothesis |

Note. Stimuli within a set were associated with each other during the study. Lesson 1 used the A, B, and C stimuli. Lesson 2 used the D, E, and F stimuli. In Lesson 3 students were required to attend to A stimuli to make decisions about how D stimuli were related to the E and F stimuli. In Lessons 2 and 3 there were three separate versions of the D stimuli, representing different types of predictions about changes in a dependent variable.

stimuli were somewhat more complex, requiring some minor modifications that are described below.

On each trial of the lessons, a sample stimulus appeared in the top box simultaneously with three comparison boxes below it. One comparison stimulus was the correct choice, one was an incorrect choice, and the third was a blank white box that also counted as an incorrect choice if selected. Positions of the comparison stimuli within the white boxes were assigned randomly for each trial.

Feedback. During training, each response resulted in accuracy feedback presented through stereo headphones and was followed by the next trial. Correct responses were followed by an ascending sound, and incorrect responses were followed by a descending sound, called “chime” and “chord,” respectively, in the Microsoft Windows XP operating system. During testing phases, clicking on any comparison stimulus immediately initiated the next trial (no feedback). During both training and testing phases, students were given visual feedback via an information box (approximately 7.6 cm by 7.6 cm) that appeared in the upper right corner of the screen. During training, the box was blue. The top half of the box stated the number of consecutive correct responses a participant needed to complete the phase (see mastery

criteria below). The bottom half displayed the number of consecutive correct responses the participant had made prior to the current trial. This counter incremented with each correct response and reset to zero with each error. During testing, the box was red. The top half of the box displayed the number of trials on the test, and the bottom showed the number of the trial on which the participant currently was working.

Mastery criterion during learning units. Each training phase was organized into two or more learning units that focused on teaching one type of relation. For example, during the training phase of Lesson 1, one unit taught the A→B relations shown in Figure 1, and another taught the C→A relations. During each learning unit, a student was considered to have demonstrated mastery after making correct responses on 12 consecutive trials. This criterion was used for two reasons. First, 12 consecutive correct responses are unlikely to occur by chance. Assuming two viable response options on each trial (ignoring the blank box) and random responding, the cumulative probability of 12 consecutive correct responses is about .0002. Second, many published studies use blocks of trials that include about six trials of each trial type (e.g., Fienup & Dixon, 2006). Thus, a criterion of 12 consecutive correct responses is

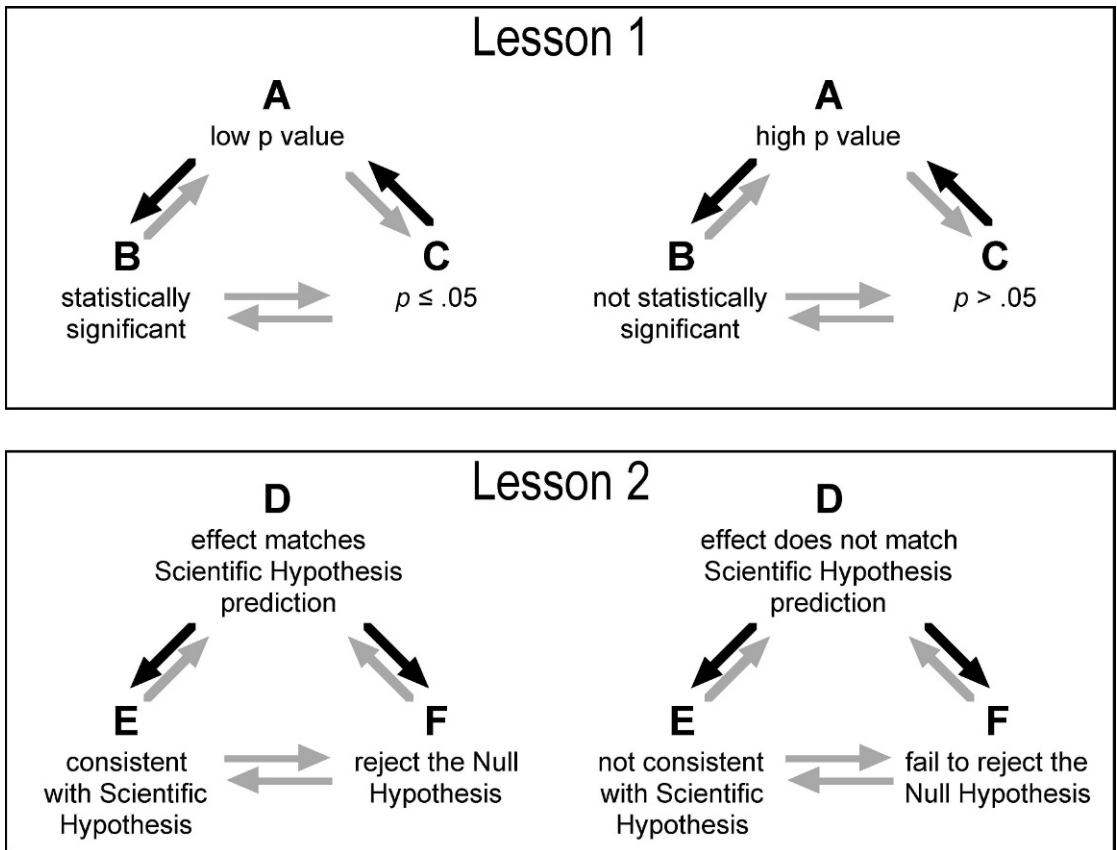


Figure 1. Summary of the relations that were taught and tested in Lessons 1 and 2. Black arrows show trained relations, and gray arrows show expected emergent relations. For exact wording contained in the stimuli, see Table 1.

consistent with common practices in this research area. Once the mastery criterion was met, a message on screen stated, "You have passed! Click this button to continue with the next learning block."

Trial sequence. Each learning unit incorporated trials reflecting two sets of stimuli, each of which could be said to "go together" according to conventions of statistical inference and the contingencies of training. For example, in the first lesson, one set of stimuli was related to *statistically significant* and the other was related to *not statistically significant*. In each consecutive pair of trials the sample and correct choice of one trial represented Set 1 and those of the other trial represented Set 2, with the sequence randomly determined within the pair. This resulted in the relation from a given stimulus set

being presented no more than twice consecutively. Trials were presented until a participant met the mastery criterion of 12 consecutive correct answers. Thus, assessments of mastery were based on either six trials from each of the two sets or seven trials from one set and five from the other set.

Mastery criterion during tests. A student who scored 89% correct or higher on a lesson's posttest was considered to have mastered that lesson and proceeded immediately to the next lesson. A student who scored lower proceeded to remediation, which consisted of repeating the lesson's learning units before taking the posttest again.

Preliminary Training

Participation began with two brief tutorials. The first familiarized students with the structure

of match-to-sample trials (a detailed description is available from the first author). The second verified that all students understood the inequality notation ($>$ and \leq) that was used in the C stimuli. In match-to-sample trials, the C stimuli ($p \leq .05$ and $p > .05$) were presented as sample stimuli and p values (drawn from .001, .01, .02, .03, .04, .05, .051, .06, .07, .08, .09, and .10) were shown as comparison stimuli. All students met the 12-trial mastery criterion within 18 trials ($M = 14$, $SD = 3.10$).

Lesson 1: A-B-C Relations (Statistical Significance)

In two learning units students learned how the following stimuli relate: p value descriptors (A stimuli), statistical significance or nonsignificance (B stimuli), and specific ranges of p values (C stimuli). Figure 1 (top) displays a pictorial representation of trained and expected emergent relations. Appendix A shows the details of the trained relations. During A \rightarrow B training, students learned to match the comparison stimuli *statistically significant* and *not statistically significant* to the samples *low p value* and *high p value*, respectively. During C \rightarrow A training they learned to match the comparison stimuli *low p value* and *high p value* to the samples $p \leq .05$ and $p > .05$, respectively.

The tests contained 48 trials, including four each of two trained A \rightarrow B and two trained C \rightarrow A relations and their untrained symmetrical variants (B \rightarrow A and A \rightarrow C) plus four each of two B \rightarrow C and C \rightarrow B emergent relations. In the latter case, on the Lesson 1 posttest, students were expected to demonstrate untaught relations between the B stimuli (*statistically significant* and *not statistically significant*) and the C stimuli (*low p value* and *high p value*, respectively).

Lesson 2: D-E-F Relations (Hypothesis Decisions in the Absence of Statistical Information)

Students learned how the following stimuli relate: scientific hypothesis paired with a description of directional effects (D stimuli), decisions regarding the scientific hypothesis (E

stimuli), and decisions regarding the null hypothesis (F stimuli). Figure 1 (bottom) displays a pictorial representation of trained and expected emergent relations. Appendix B shows the details of the trained relations and the sequence in which they were taught.

Two aspects of the stimuli bear special explanation. The first key point is that, although all D stimuli included a scientific hypothesis and a description of directional research outcomes, there were three different kinds of D stimuli. On a given trial, the scientific hypothesis predicted an increase in the dependent variable, a decrease in the dependent variable, or a change in the dependent variable. Thus, the training incorporated both one-tailed and two-tailed scientific hypotheses. The direction of effect with which the scientific hypothesis was paired either matched the prediction (e.g., *the dependent variable increased*) or failed to match it (e.g., *the dependent variable did not increase*, which could imply a decrease or no change).

The second key feature of the stimuli is that traditional statistical language was altered to make the lesson consistent with instruction in an introductory psychology statistics course at the university at which the study was conducted. Typically, statisticians refer to rejecting or failing to reject hypotheses, both scientific and null (Huck, 2000). For study purposes, this terminology was applied to statements involving the null hypothesis (F stimuli). When the scientific hypothesis was concerned (E stimuli), however, *not consistent with the scientific hypothesis* replaced *reject*, and *consistent with the scientific hypothesis* replaced *fail to reject*. During D \rightarrow E training, students learned to match comparison stimuli *consistent with the scientific hypothesis* and *not consistent with the scientific hypothesis* with samples in which an effect did or did not, respectively, descriptively match the scientific hypothesis prediction. There were three D \rightarrow E learning units, one in which the dependent variable was predicted to increase,

one in which it was predicted to decrease, and one in which it was predicted to change. During D→F training, students learned to match comparison stimuli *reject the null hypothesis* and *fail to reject the null hypothesis* with samples in which an effect did or did not, respectively, descriptively match the scientific hypothesis prediction. There were three D→F learning units, one in which the dependent variable was predicted to increase, one in which it was predicted to decrease, and one in which it was predicted to change.

Learning units were organized into six units according to the direction of scientific hypothesis prediction. Students first completed both D→E and D→F training involving sample stimuli in which the dependent variable was predicted to increase. They next completed both D→E and D→F training involving sample stimuli in which the dependent variable was predicted to decrease, followed by both D→E and D→F training involving sample stimuli in which the dependent variable was predicted to change.

The tests contained 56 trials, including two trials each of 12 trained D→E (six from each of the two sets) and 12 trained D→F (six from each of the two sets) relations plus their untrained symmetrical variants (E→D and F→D), plus two E→F and two F→E emergent relations. In the latter case, on the Lesson 2 posttest, students were expected to demonstrate untaught relations between the E stimuli (*consistent with the scientific hypothesis* and *not consistent with the scientific hypothesis*) and the F stimuli (*reject the null hypothesis* and *fail to reject the null hypothesis*, respectively). Thus, through emergent relations, students were expected to understand that decisions about the null and scientific hypotheses are mutually exclusive.

Lesson 3: Contextual Relations

To promote control of responding by both hypothesis information and statistical information, in Lesson 3 the sample stimuli paired the D stimuli of Lesson 2 (scientific hypothesis plus

a description of the direction of an effect) with the A stimuli of Lesson 1 (*low p value* and *high p value*). The D and A stimuli were displayed next to each other in one sample-stimulus box (15.24 cm wide). Students were instructed to “use both pieces of information to make decisions.” Figure 2 (top) displays a pictorial representation of the relations involved in Lesson 3.

The A stimuli served a function analogous to that of the nationality or profession cues in the contextual control example in the introduction: They signaled the function of D stimuli that indicated a match between the direction of effect and the scientific hypothesis prediction. The contextual cue *low p value* (A) signaled that this D stimulus should be matched with *consistent with the scientific hypothesis* (E) and *reject the null hypothesis* (F). The contextual cue *high p value* (A) signaled that this D stimulus should be matched with *not consistent with the scientific hypothesis* (E) and *fail to reject the null hypothesis* (F). It should be noted that the function of the contextual cue was itself conditional under the contingencies of Lesson 3 training, which is a more complex arrangement than illustrated in the example in the Introduction. When the D stimulus indicated a mismatch between the direction of an effect and the scientific hypothesis prediction, contextual control by the A stimuli did not apply. Correct responding always matched the D stimulus with *not consistent with the scientific hypothesis* (E) and *fail to reject the null hypothesis* (F). Speaking colloquially, if an effect runs counter to the scientific hypothesis, then the statistical significance of that effect is irrelevant to hypothesis decisions.

There were 12 learning units. Appendix C specifies the relations that were taught and in what sequence. Across four consecutive learning units, the scientific hypothesis of the D stimulus remained constant, predicting an increase, decrease, or change, respectively, in the dependent variable, as was the case in Lesson 2.

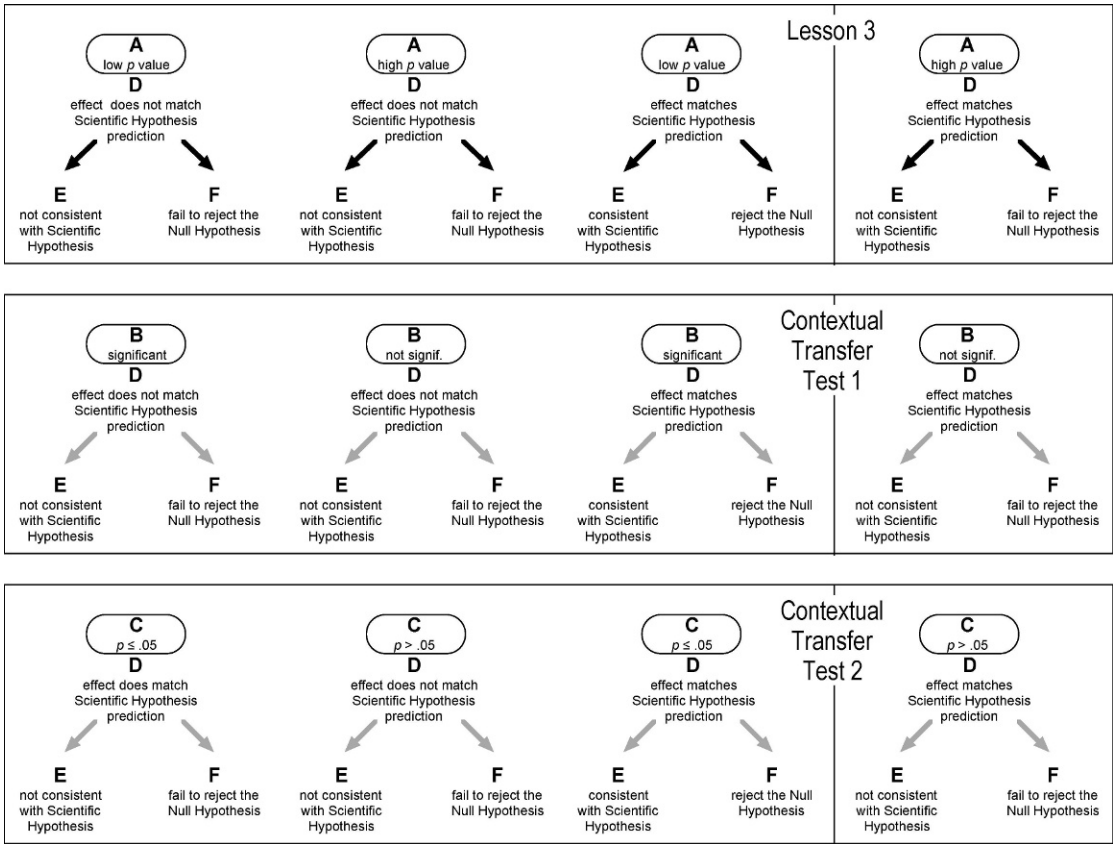


Figure 2. Summary of the relations that were taught and tested in Lesson 3 and the subsequent contextual transfer tests. Black arrows show trained relations, and gray arrows show expected emergent relations. The contextual transfer tests were identical to Lesson 3 tests except that new Lesson 1 stimuli were substituted for the A stimulus. See text for details. For exact wording contained in the stimuli, see Table 1. For all panels, relations located to the right of the vertical line are described in text as modified relations; others are described as unmodified relations.

Within each set of four learning units, the E stimuli (*consistent with the scientific hypothesis* and *not consistent with the scientific hypothesis*) served as comparisons for the first pair, and the F stimuli (*reject the null hypothesis* and *fail to reject the null hypothesis*) served as comparisons for the last pair. Within each pair of learning units, the A stimulus was *low p value* for the first unit and *high p value* for the second unit. In this way training incorporated all possible combinations of D and A sample-stimulus compounds with E and F comparison stimuli.

Lesson 3 tests contained 48 trials, four each of the 12 relations that were involved in training. Thus, unlike in Lessons 1 and 2, the

pretest and posttest evaluated only relations that were taught explicitly, although emergent relations were expected. Because Lesson 3 employed a stimulus (A) from Lesson 1 as a contextual cue, it was expected that other Lesson 1 stimuli (B and C) would come to fulfill the contextual function without any explicit training. This possibility was evaluated in two contextual transfer tests. It also was possible that Lesson 3 training could have adverse effects on the equivalence class that emerged in Lesson 1. This possibility was evaluated through three maintenance tests. Unlike previous tests, the contextual transfer and maintenance tests were administered only

once, with no mastery criterion. After completion of these tests, students were dismissed from the study.

Contextual transfer tests. If the contextual training of Lesson 3 met its academic goals, then the contextual function served by the A stimulus also could be served by stimuli from Lesson 1 that had not been part of contextual training but that had been shown previously to be equivalent with A. Two contextual transfer tests evaluated this interchangeability. Contextual Transfer Test 1 was identical to the contextual posttest, except that B stimuli were substituted for A stimuli in the samples. Specifically, *statistically significant* replaced *low p value* and *not statistically significant* replaced *high p value*. Contextual Transfer Test 2 substituted C ($p \leq .05$ and $p > .05$, respectively) stimuli for the A stimuli. The middle and bottom panels of Figure 2 display pictorial representations of the relations involved in these two tests.

Maintenance of A-B-C relations. The contextual training of Lesson 3 may be thought of as using Lesson 1 information to redefine some of the relations that had been mastered in Lesson 2. Past research has shown that when previously learned relations are redefined in ways that are partly analogous to the procedures of the present contextual training, one potential outcome is the disruption of equivalence classes (e.g., Pilgrim & Galizio, 1995). It was possible, therefore, that the modification of Lesson 2 relations would disrupt the Lesson 1 stimulus class with which contextual training associated it, resulting in no systematic relations among the A-B-C stimuli. Another possibility is that, just as the A stimuli cued a "reassignment" of D stimuli between equivalence classes, so too might D stimuli promote a reassignment of function for the A stimulus. The result would be systematic relations among A-B-C stimuli that were different from the ones promoted during Lesson 1 (and therefore academically inappropriate).

To illustrate the latter problem, consider the case in Lesson 2 in which results corresponding to a scientific hypothesis prediction (D stimulus) were presented along with the options *consistent with the scientific hypothesis* and *not consistent with the scientific hypothesis*, with the former as the correct choice. In Lesson 3, the same stimulus paired with *high p value* (A) now prompted a match with *inconsistent with the scientific hypothesis* (a reversal of function). Now consider the role of the A stimuli in Lesson 1. Given *low p value* and the options *statistically significant* and *not statistically significant*, the former would be the correct choice. What might happen if the same A stimulus were presented with the D stimulus mentioned above following Lesson 3, along with the options *statistically significant* and *not statistically significant*? If D modulated the function of A, a student might choose the latter.

The purpose of maintenance tests was to evaluate whether Lesson 1 relations survived as originally trained despite the pairing of stimuli from different lessons in a compound sample stimulus. During these tests, the sample stimuli mirrored those of Lesson 3 in pairing a stimulus from Lesson 1 (A, B, or C) with a stimulus from Lesson 2 (D). However, maintenance tests differed from Lesson 3 tests in that the comparison stimuli were the A-B-C stimuli instead of E and F stimuli. This distinction rendered the D stimulus in the sample irrelevant to selecting a comparison stimulus. For example, whether $p \leq .05$ qualifies as *statistically significant* is unrelated to whether results match the predictions of a scientific hypothesis. If contextual training succeeded as intended, students would continue to match Lesson 1 stimuli to one another without distraction from the portion of the complex sample stimulus that was drawn from Lesson 2.

There were three maintenance tests. Figure 3 displays pictorial representations of each of the tests. Each test contained 48 trials, including two each of 24 trial types derived by combining

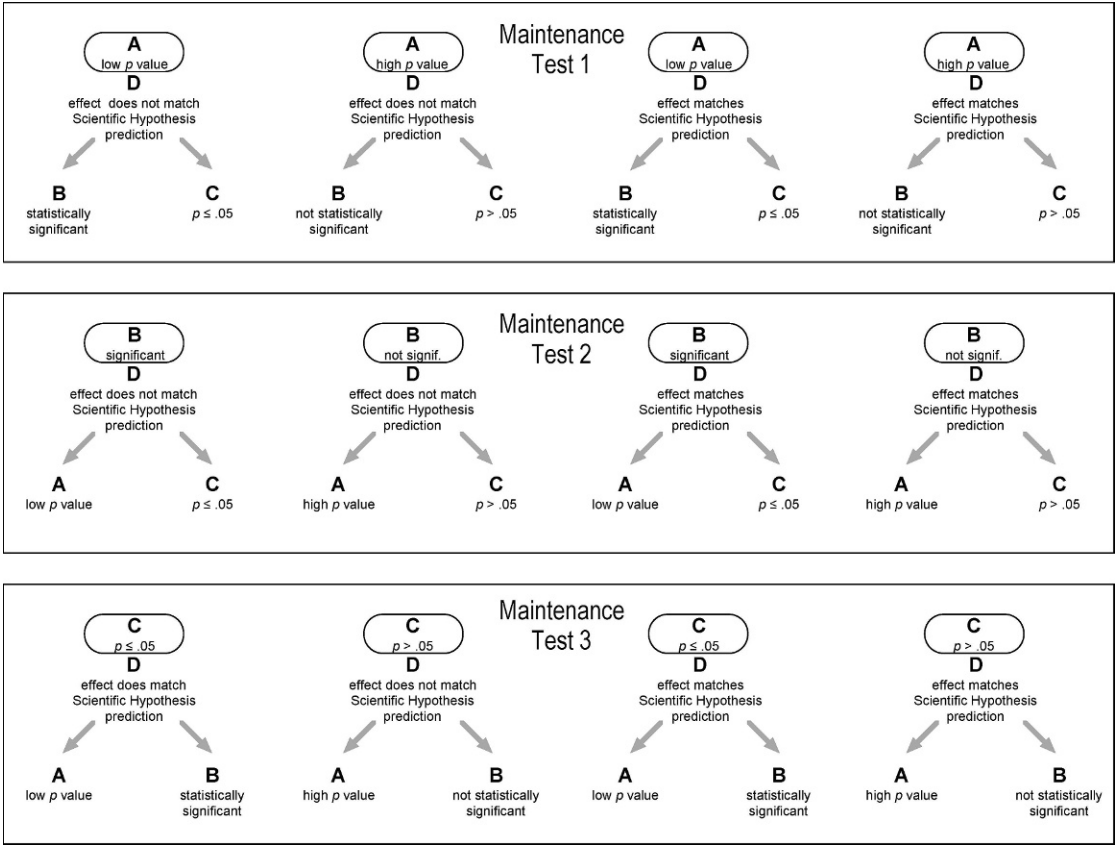


Figure 3. Summary of the relations that were assessed in A-B-C maintenance tests. Arrows show relations trained during Lesson 1. See text for details.

two Lesson 1 sample stimuli (e.g., in Test 1, the A stimuli *low p value* and *high p value*), two D stimuli (in which the effect matched or did not match the prediction) in three variants each (predicting independent variable increase, decrease, and change), and two types of Lesson 1 stimuli as comparisons (e.g., the B and C stimuli in Test 1). In Maintenance Test 1, the sample compounds were identical to those of Lesson 3 training (D and A stimuli); on different trials, the comparison stimuli were either the B stimuli (*statistically significant* and *not statistically significant*) or the C stimuli ($p \leq .05$ and $p > .05$). In Maintenance Test 2, the sample compound combined D and B stimuli as in Contextual Transfer Test 1, and the comparison stimuli were either A or C stimuli. In Maintenance Test 3, the sample compound

combined D and C stimuli as in Contextual Transfer Test 2, and the comparison stimuli were either A or B stimuli.

RESULTS

Overview

Table 2 provides an overview of the results of the computerized lessons, comparing pretest and posttest scores of each of the three lessons. In all cases, students made errors on a considerable percentage of pretest trials prior to training. Following training, scores improved to near 100% on the posttest. Thus, the computerized lessons created mastery of skills relevant to statistical inference (Lesson 1), hypotheses decisions (Lesson 2), and the contextually controlled conjunction of these

Table 2
Overall Scores on Pretest and Posttests

| Participant | Lesson 1: A-B-C | | Lesson 2: D-E-F | | Lesson 3: Contextual | |
|-------------------|-----------------|----------|-----------------|----------|----------------------|----------|
| | Pretest | Posttest | Pretest | Posttest | Pretest | Posttest |
| 1 | 33 | 98 | 66 | 96 | 63 | 100 |
| 2 | 35 | 65 (100) | 41 | 100 | 48 | 100 |
| 3 | 33 | 98 | 36 | 98 | 73 | 100 |
| 4 | 69 | 100 | 45 | 100 | 75 | 81 (100) |
| 5 | 35 | 100 | 43 | 96 | 75 | 100 |
| 6 | 33 | 100 | 39 | 98 | 46 | 100 |
| 7 | 31 | 100 | 54 | 100 | 73 | 100 |
| 8 | 60 | 98 | 70 ^a | 95 | 71 | 100 |
| 9 | 35 | 98 | 45 | 56 (100) | 69 | 100 |
| 10 | 31 | 100 | 45 | 100 | 50 | 96 |
| Mean ^b | 40 | 99 | 48 | 98 | 64 | >99 |

Note. Numbers in parentheses are scores from a second attempt at a posttest following training remediation that was prompted by a failed initial attempt at the posttest.

^a 69.6% (less than exclusion criterion).

^b For students who were tested twice, only the second test was considered in the mean.

two repertoires (Lesson 3) for every student. The following sections provide a more detailed inspection of these outcomes.

Lesson 1: A-B-C Relations (Statistical Significance)

Overall, this lesson improved student scores from a mean of 40% correct ($SD = 13.4$) on the pretest to a mean of 99% correct ($SD = 1.0$) on the posttest. Figure 4 shows Lesson 1 pretest accuracy by type of relation. Here the relations are grouped thematically as a teacher might organize them. In the left two columns, results for relations that were taught directly (e.g., $C \rightarrow A$) are pooled with those of symmetrical variants that were not taught directly (e.g., $A \rightarrow C$). Thus, the display departs from convention in the stimulus equivalence literature, which typically treats symmetrical variants of trained relations as separate emergent skills. We will address symmetrical relations further in the Discussion; for present purposes, Figure 4 adopts the practical perspective of considering jointly the relations that involve the same pair of stimuli.

Figure 4 (top left) shows that most students did not reliably match *low p value* with *statistically significant* and *high p value* with *not statistically significant* ($A \rightarrow B$, $B \rightarrow A$) prior to

training. Scores lower than 50% correct suggest that most students entered the study with a bias for associating *statistically significant* with *high p value*. Figure 4 (top middle) shows that most of the students were able to match *low p value* with $p \leq .05$ and *high p value* with $p > .05$ ($C \rightarrow A$, $A \rightarrow C$) prior to any training. Most students scored poorly on the transitive relations that were expected to emerge untaught from Lesson 1 training.

During Lesson 1 training (see Table 3), the students required a total of 25 to 62 trials to meet the mastery criterion of 12 consecutive correct responses in each of two learning units; the upper extreme reflects the fact that Student 2 failed an initial attempt at the Lesson 1 posttest and had to complete training twice. Following training, students completed the Lesson 1 posttest almost without error, and accuracy for emergent transitive relations ($B \rightarrow C$, $C \rightarrow B$) paralleled that for the directly taught relations.

Lesson 2: D-E-F Relations (Hypothesis Decisions in the Absence of Statistical Information)

Overall, this lesson improved student scores from a mean of 46% correct ($SD = 9.0$) on the pretest to a mean of about 98% correct ($SD =$

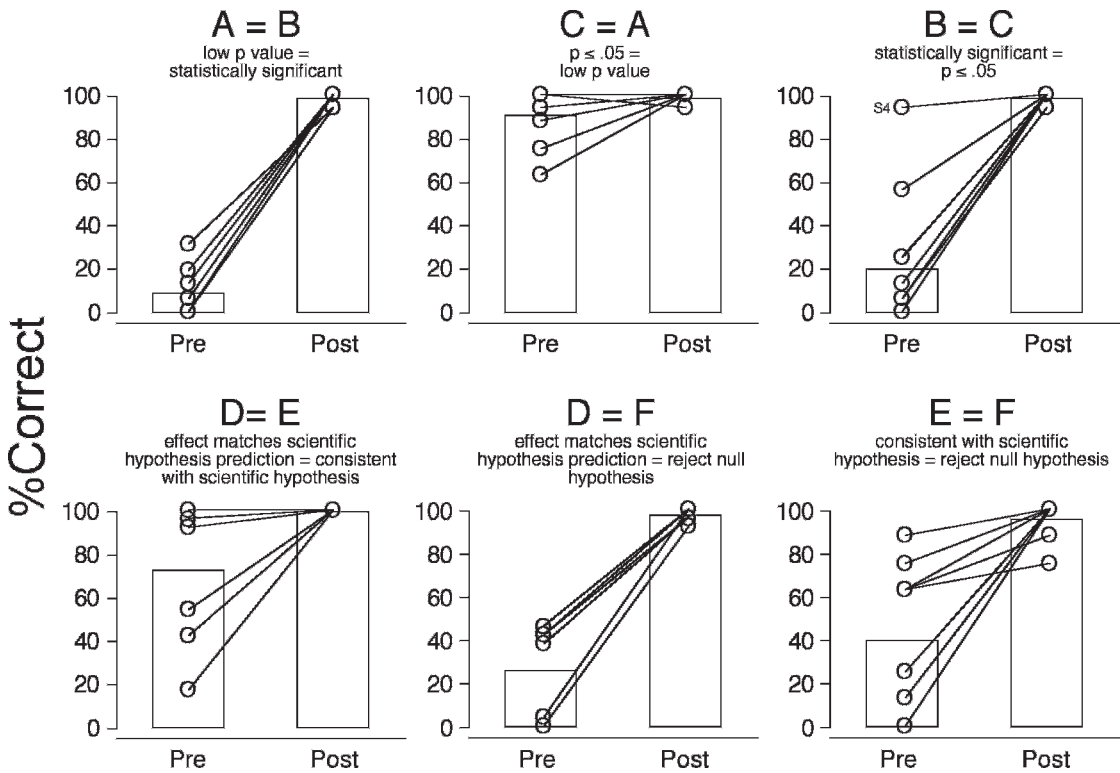


Figure 4. Summary of pretest and posttest results for Lessons 1 (top) and 2 (bottom). In each panel, bars show the average accuracy of 10 students, and circles show results of individual students. The label above each panel shows an example of the stimuli (see Table 1 for exact wording) that correct responding matched; each relation type included stimuli from two classes. Thus, C-A relations involved both matching $p \leq .05$ with low p value (shown) and $p > .05$ with high p value (not shown). Results were pooled for each trained relation type (e.g., C→A) and its symmetrical variant (e.g., A→C). See text for explanation of data aggregation.

2.0) on the posttest. Figure 4 (bottom row) shows Lesson 2 pretest accuracy by type of relation. Figure 4 (bottom middle) shows that most students did not match combinations of scientific hypotheses and results to decisions regarding the null hypothesis (D→F, F→D) reliably prior to training. Accuracy was well below chance (50%) for several students; thus, prior to training, students often generated inappropriate matches, such as *inconsistent with the scientific hypothesis* with *reject the null hypothesis*. Figure 4 (bottom left) shows that half the students were able to match various combinations of scientific hypotheses and results to decisions regarding the scientific hypothesis (D→E, E→D). For 8 of 10 students, pretest scores were $\leq 70\%$ accurate

for transitive relations (F→F, F→E; bottom right) that were expected to emerge untaught from Lesson 2 training.

During Lesson 2 training (see Table 3), the students required a total of between 72 and 159 trials to reach mastery criterion; the upper extreme reflects the fact that Student 9 failed an initial attempt at the Lesson 2 posttest and had to complete training twice. Following training, all students completed the posttest at a high level of proficiency, and accuracy for the emergent transitive relations (E→F, F→E) paralleled that of the directly taught relations.

Lesson 3: Contextual Relations

Scores improved for all students between the Lesson 3 pretest ($M = 64\%$, $SD = 11.8$) and

Table 3
Trials to Mastery Criterion for Lesson 1 (A-B-C) and Lesson 2 (D-E-F) Training

| Participant | Lesson 1 | | Lesson 2 | | | | | |
|-------------------|----------|---------|-------------|---------|-------------|---------|-----------|---------|
| | A→B | C→A | DV increase | | DV decrease | | DV change | |
| | | | D→E | D→F | D→E | D→E | D→F | D→E |
| 1 | 14 | 13 | 12 | 17 | 12 | 12 | 19 | 12 |
| 2 | 23 (12) | 13 (14) | 12 | 13 | 12 | 12 | 12 | 12 |
| 3 | 13 | 19 | 12 | 12 | 13 | 12 | 12 | 12 |
| 4 | 13 | 12 | 12 | 13 | 12 | 12 | 12 | 12 |
| 5 | 13 | 12 | 12 | 13 | 12 | 12 | 12 | 12 |
| 6 | 14 | 12 | 12 | 12 | 12 | 12 | 12 | 12 |
| 7 | 19 | 20 | 12 | 25 | 26 | 12 | 12 | 12 |
| 8 | 27 | 12 | 12 | 13 | 12 | 20 | 12 | 12 |
| 9 | 13 | 12 | 12 (12) | 14 (14) | 21 (12) | 12 (12) | 12 (12) | 14 (12) |
| 10 | 13 | 13 | 14 | 18 | 12 | 12 | 12 | 12 |
| Mean ^a | 17 | 15 | 13 | 16 | 16 | 14 | 14 | 13 |

Note. Number of trials required to meet the mastery criterion of 12 consecutive correct responses in each training block of Lessons 1 and 2. Number in parentheses represent remediation that was required after failure of a posttest; see text for details. In Lesson 2 columns, major headers refer to the direction of effect in a dependent variable (DV) that was predicted by the scientific hypothesis in D stimuli.

^a For students who required remedial training, trials combined from both the initial and remedial training were considered in the mean.

posttest ($M = 100\%$, $SD = 1.2$). Yet overall scores obscure the fact that many of the relations that were part of Lesson 3 could be answered correctly on the basis of Lesson 2 training without reference to the inferential statistics information that was part of Lesson 3 sample stimuli. In some cases, the statistical information is not required to render hypothesis decisions (e.g., given an effect in a different direction than predicted by the scientific hypothesis, select *not consistent with the scientific hypothesis*, regardless of the associated p value). In other cases, the statistical information corroborates a qualitative appraisal of results (e.g., given an effect in the direction predicted by the scientific hypothesis and a low p value, select *consistent with the scientific hypothesis*). In cases like these, a student who completed only the present Lesson 2 and ignored statistical information on the Lesson 3 tests would be expected to answer correctly. The relevant relations will be referred to as having been left unmodified by Lesson 3; they comprised 75% of trials on the Lesson 3 pretest and posttest. For the remaining relations, correct responding

required attention to both the hypothesis results match (D stimulus) and statistical information (A stimulus). For example, given an effect in the direction predicted by the scientific hypothesis and a *high p value*, a correct response would be *inconsistent with the scientific hypothesis* (E stimulus). In such cases, a student who responded strictly in accordance with Lesson 2 training would be expected to respond incorrectly. The relevant relations will be referred to as having been modified by Lesson 3; they comprised 25% of trials on the Lesson 3 pretest and posttest.

Figure 5 summarizes Lesson 3 pretest scores. As expected based on the preceding analysis, for six students pretest scores were near 100% correct for unmodified relations (left panel) and 0% correct for modified relations (right panel). Thus, these students began training with a tendency to make hypothesis decisions based solely on the correspondence between hypotheses and descriptive results, just as they were taught in Lesson 2. For Students 1, 2, 6, and 10, pretest scores were near chance (50% correct) for both unmodified and modified

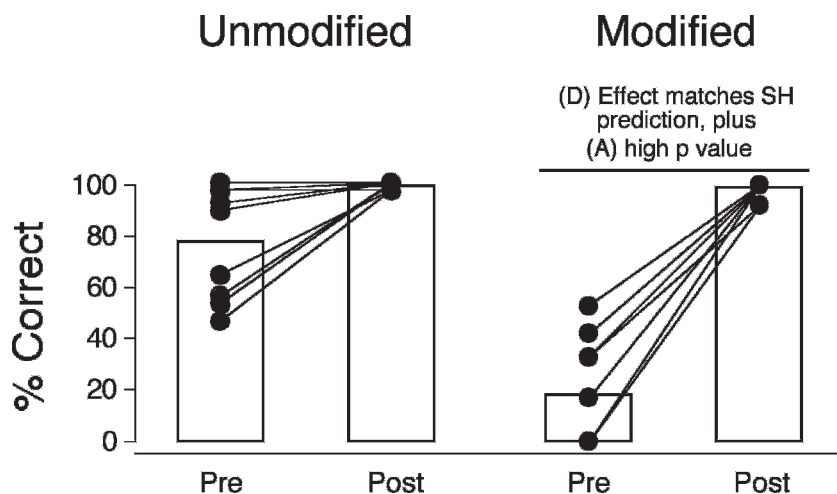


Figure 5. Summary of pretest and posttest results for Lesson 3, reported separately for relations in which the function of the D stimulus was unmodified (left) or modified (right) during Lesson 3 training. The label above the right panel shows the stimuli that were involved in modified relations (see Table 1 for exact wording). In each panel, bars show the average accuracy of 10 students, and circles show results of individual students.

relations, indicating that the presence of statistical information adversely affected the D→E and D→F relations that had been taught in Lesson 2.

During Lesson 3 training (see Table 4), to meet the mastery criterion of 12 consecutive correct responses in each of six learning units, the students required a total of 160 to 313 trials; the upper extreme reflects the fact that Student 4 failed an initial attempt at the Lesson 3 posttest and had to complete training twice. Figure 5 summarizes the Lesson 3 posttest results. On modified relations, only Student 10 (92% correct) made errors on the posttest; thus, Lesson 3 training succeeded in redefining selected D→E and D→F relations of Lesson 2 on the basis of statistical information. On unmodified relations, students with high pretest accuracy also did well on the posttest; this portion of the Lesson 2 repertoire was not adversely affected by Lesson 3 training. For students with low pretest scores on unmodified relations, accuracy increased to near 100% on the posttest.

Contextual transfer tests. The Lesson 3 posttest did not incorporate any emergent relations, although emergent relations were expected to

result from Lesson 3 training to the extent that B and C stimuli of Lesson 1 were indeed equivalent to (interchangeable with) the A stimuli that were part of Lesson 3 sample stimuli. Figure 6 summarizes the results of the two contextual transfer tests. The left column of panels shows accuracy when B (statistical significance, Test 1; top left panel) or C (p compared to $\alpha = .05$, Test 2; bottom left panel) stimuli were substituted for the A stimulus (low or high p) in the sample stimuli of unmodified relations. Accuracy was $\geq 89\%$ on both tests, with the exception of Student 8 (72%) on Contextual Transfer Test 1. On this test, for unknown reasons, Student 8 made errors on 13 of the final 14 trials (not specific to any particular relation type), but responded virtually without error on earlier portions of Test 1 and on all of the subsequent Contextual Transfer Test 2.

Figure 6 (right column) shows analogous outcomes for modified relations. In Test 2 (bottom right), when an effect corresponded to the scientific hypothesis prediction but $p > .05$ (C) was substituted for *high p value* (A), 9 of 10 students almost always made correct hypothesis decisions. The lone exception was Student 4,

Table 4
Trials to Mastery Criterion for Lesson 3

| D Sample Comparison A Sample Participant | DV increase | | | | DV decrease | | | | DV change | | | |
|---|-----------------|------------------|-----------------|------------------|-----------------|------------------|-----------------|------------------|-----------------|------------------|-----------------|------------------|
| | Scientific (E) | | Null (F) | | Scientific (E) | | Null (F) | | Scientific (E) | | Null (F) | |
| | Low <i>p</i> | High <i>p</i> | Low <i>p</i> | High <i>p</i> | Low <i>p</i> | High <i>p</i> | Low <i>p</i> | High <i>p</i> | Low <i>p</i> | High <i>p</i> | Low <i>p</i> | High <i>p</i> |
| 1 | 12 | 20 | 24 | 12 | 12 | 12 | 12 | 12 | 15 | 12 | 12 | 12 |
| 2 | 14 | 22 | 14 | 12 | 12 | 12 | 12 | 12 | 20 | 12 | 12 | 12 |
| 3 | 12 | 21 | 14 | 12 | 12 | 12 | 12 | 12 | 15 | 12 | 12 | 14 |
| 4 | 12 (12) | 25 (13) | 14 (12) | 15 (12) | 12 (12) | 12 (12) | 12 (12) | 13 (12) | 12 (12) | 16 (12) | 12 (12) | 13 (12) |
| 5 | 12 | 18 | 12 | 12 | 12 | 13 | 12 | 23 | 12 | 12 | 12 | 22 |
| 6 | 12 | 21 | 12 | 12 | 12 | 12 | 12 | 12 | 12 | 12 | 21 | 12 |
| 7 | 12 | 22 | 12 | 22 | 16 | 12 | 14 | 12 | 12 | 12 | 12 | 12 |
| 8 | 12 | 23 | 13 | 12 | 24 | 12 | 12 | 12 | 12 | 12 | 12 | 12 |
| 9 | 12 | 16 | 25 | 12 | 12 | 12 | 12 | 12 | 20 | 12 | 12 | 12 |
| 10 | 12 | 38 | 12 | 12 | 12 | 12 | 12 | 12 | 12 | 12 | 12 | 12 |
| Mean ^a | 13 | 24 | 16 | 15 | 15 | 13 | 13 | 14 | 15 | 14 | 14 | 15 |

Note. Number of trials required to meet the mastery criterion of 12 consecutive correct responses in each training block of Lesson 3. Number in parentheses represent remediation that was required after failure of a posttest; see text for details. Column headers are organized as follows. First-level headers refer to the direction of a dependent variable (DV) effect that was predicted by the Scientific hypothesis in the D stimuli. Third-level headers refer to the A (contextual) stimulus with which the D stimulus was presented. Second-level headers refer to the type of comparison stimuli from which students chose.

^a For students who required remedial training, trials combined from both the initial and remedial training were considered in the mean.

who routinely ignored statistical information in making hypothesis decisions. In Test 1 (top right), when an effect corresponded to the scientific hypothesis prediction but *not statistically significant* (B) was substituted for *high p value* (A), 7 of 10 students usually made correct hypothesis decisions. Of the remaining participants, only Student 4 (who responded as in Transfer Test 1 by ignoring statistical information consistently) showed an unequivocal failure of contextual control transfer. Student 8 showed a relation-nonspecific pattern of errors that was described above. Like Student 8, Student 3's errors all occurred sequentially on Transfer Test 1, but they were relation specific, including the first six trials of modified relations. Thereafter, Student 3 responded correctly to modified relations on the remainder of Transfer Test 1 and all of Transfer Test 2. Assuming that the A-B-C stimuli remained equivalent for this student, the pattern is reminiscent of the phenomenon known as delayed emergence, in which untrained relations strengthen with

testing experience in the absence of feedback (e.g., Sidman, Kirk, & Willson-Morris, 1985). Sidman (1994) has suggested that testing provides an opportunity to explore competing hypotheses about what stimuli "go together" and to determine which response strategies map most consistently onto the various configurations of stimuli (i.e., test questions) that are part of the tests. It is possible that with more extended exposure to the tests, Student 3 might have demonstrated mastery of TCC relations.

Maintenance of A-B-C relations. The purpose of these tests was to determine whether the presence of the D stimulus (from Lesson 2) would have adverse effects on the maintenance of relations among the A-B-C stimuli. Recall that the trials were structured so that a correct response was defined by matching two Lesson 1 stimuli when the sample included a Lesson 2 (D) stimulus. Median accuracy on the three tests combined was 90%. When the D stimulus described results that matched the scientific hypothesis prediction, students usually respond-

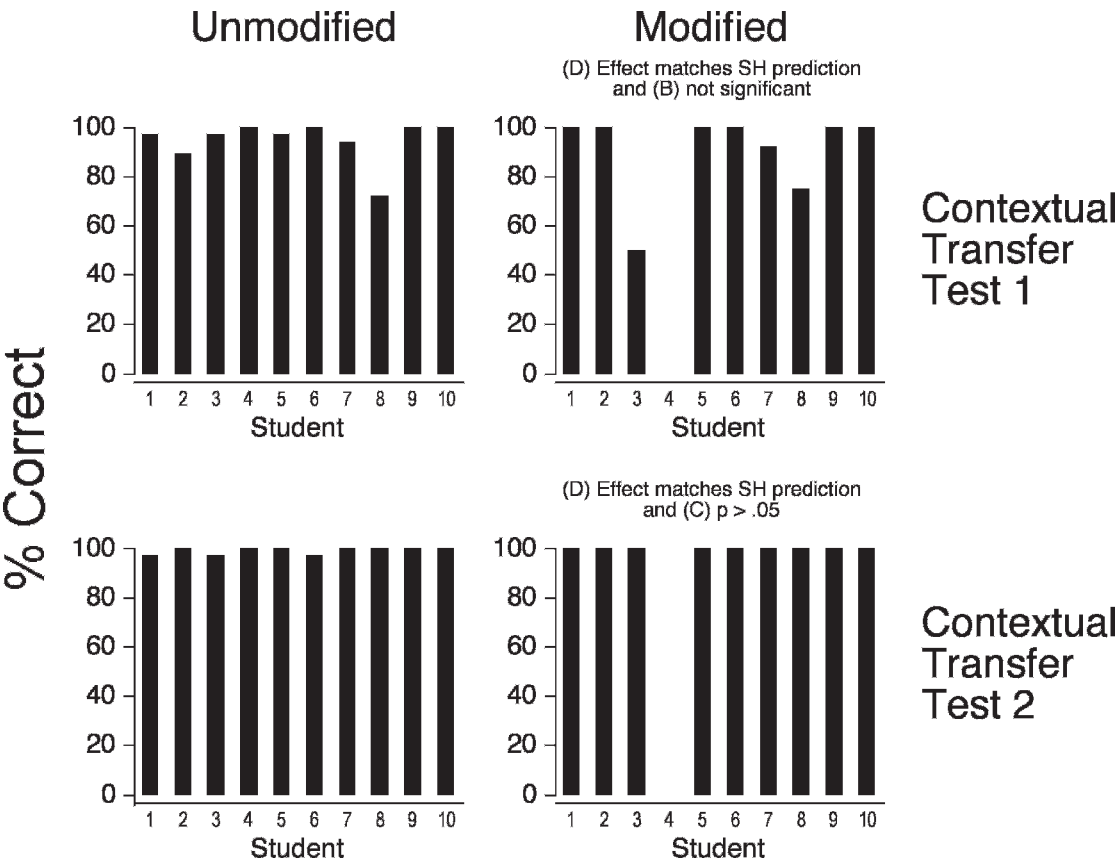


Figure 6. Summary of results for the contextual transfer tests, reported separately for relations in which the function of the D stimulus was unmodified (left) or modified (right) during Lesson 3 training. The label above the right column panels show the stimuli that were involved in modified relations (see Table 1 for exact wording). In each panel, bars show the accuracy of individual students.

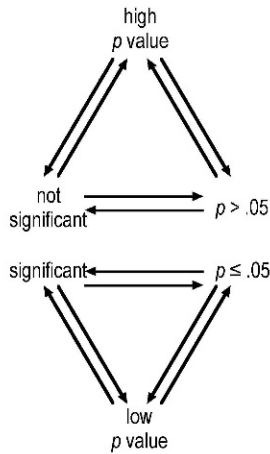
ed consistent with A-B-C relations that were established during Lesson 1. Median accuracy on these relations was 94%, and no systematic error patterns were detected. For Students 1, 4, 7, 8, and 9, this was true as well when the D stimulus described results that did not match the scientific hypothesis prediction (median accuracy = 97%). Figure 7 (top left) summarizes this outcome as black lines indicating the relations that are expected based on Lesson 1 training and testing. For each relevant relation, the line starts at the sample stimulus and culminates in an arrow designating the correct comparison stimulus.

The top portion of each remaining panel in Figure 7 shows that when a “does not match”

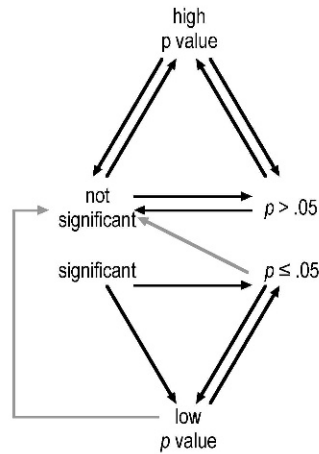
D stimulus was paired with an A, B, or C stimulus indicating nonsignificant results, Students 2, 3, 5, 6, and 10 usually responded as they had in Lesson 1. For these students, systematic error patterns occurred only when a “does not match” D stimulus was paired with an A, B, or C stimulus describing a statistically significant outcome (bottom portion of each panel). Figure 7 summarizes these error patterns as gray lines. For each relevant relation, the line starts at the sample stimulus and culminates in an arrow designating the incorrect comparison stimulus that a student frequently chose.

For Student 3, when significance was expressed in the sample as *low p value* or $p \leq .05$, the comparison *not significant* was chosen.

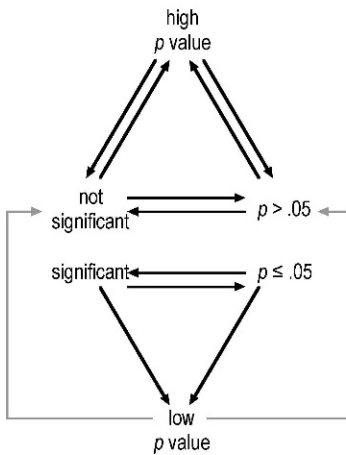
Effect Does Not Match S.H. Prediction



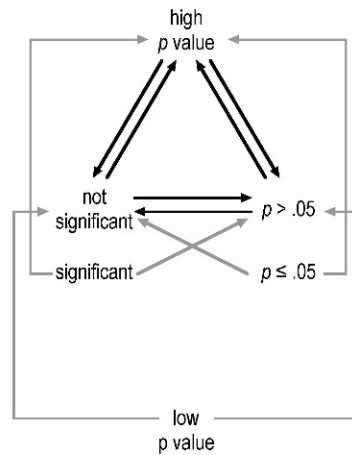
Students 1, 4, 7, 8, 9
consistent with Lesson 1



Student 3



Students 5, 6, 10



Student 2

Figure 7. A-B-C maintenance tests: dominant response patterns for relations in which the sample contained a D stimulus indicating that research results did not correspond to predictions of the scientific hypothesis. Stimuli are those described in Table 1 and in Figure 3. Each arrow summarizes one tested relation. Arrows originate with the stimulus that accompanied the D stimulus in the sample and end with a student's preferred comparison stimulus (thus, sample \rightarrow comparison). Black arrows show responding that was consistent with Lesson 1 training. Gray arrows show responding in which Lesson 1 class membership was disrupted. See text for additional explanation.

For Students 5, 6, and 10, when significance was expressed in the sample as *low p value*, *not significant* was chosen. Thus, for these students, Lesson 1 relations were preserved except in some instances in which the B or C stimuli served as comparisons and then only when a

statistically reliable effect did not match the scientific hypothesis prediction. Interestingly, the symmetrical version of the problem relations (in which B or C stimuli were part of the sample) remained intact. For Student 2, regardless of how significance was expressed in

the sample, comparisons indicating nonsignificance were chosen.

Overall, the contextual training of Lesson 3, although successful in generating correct hypothesis decisions (contextual test and contextual transfer tests), apparently had the side effect of confusing some students about the relations among stimuli indicating statistical significance. To be clear, however, systematic errors occurred on maintenance tests only for some students, and only in cases in which an effect that did not match the scientific hypothesis achieved statistical significance. Interestingly, these errors in labeling an effect as significant or nonsignificant usually were uncorrelated with other problems regarding statistical significance. Students who exhibited consistent errors on the maintenance tests did not make systematic errors on the trained contextual relations of Lesson 3 or, with one exception (Student 3, Figure 6), on the emergent contextual relations of the contextual transfer tests.

DISCUSSION

Instructional Success and Efficiency

The programmed lessons addressed the difficult (e.g., Kranzler, 2007) challenge of introducing students to preliminary concepts of statistical inference. The results corroborate those of a pilot investigation (Fienup et al., 2009) in showing that conditional reasoning in statistically informed hypothesis decisions can be established using procedures based on stimulus equivalence. One way to evaluate the success of this instruction is in the context of letter grades that often are awarded in academic systems. Based on a commonly used scale in which letter grades are separated by 10% of accuracy (90% = A, 80% = B, etc.), all students in the present study would have earned an A on the Lesson 1 and 2 posttests. Considering all trials of the Lesson 3 contextual posttest and the closely related contextual transfer tests, 9 of 10 students would have earned an A, with Student 4 (88% correct) earning a high B. Finally, even on the maintenance tests, which revealed isolated

difficulties for some individuals, 7 of 10 students would have earned an overall grade of A, with Students 2, 3, and 4 earning respectable marks (77%, 87%, and 79%, respectively). Moreover, these successes were achieved with relatively little investment of student time. Not considering the time devoted to informed consent, transitions between activities, and other nonacademic activities, all students completed the study in less than 90 min, with a large portion of that time devoted to assessment (pretests and posttests). The instructional (training) phases of the study were brief, lasting about 15 min or less for each student.

During each of the three lessons, the students readily learned what they were taught about inferential statistics and hypothesis decision making: two $A \rightarrow B$ relations and two $C \rightarrow A$ relations in Lesson 1, six $D \rightarrow E$ relations and six $D \rightarrow F$ relations in Lesson 2, and 24 relations involving combinations of D and A stimuli as samples and either E or F stimuli as comparisons in Lesson 3. As a result, they also demonstrated several emergent or untaught relations. These included the 40 symmetrical variants of relations that were taught directly, plus transitive relations among stimuli that were never paired directly during the learning units: two $B \rightarrow C$ and two $C \rightarrow B$ relations in Lesson 1, two $E \rightarrow F$ and two $F \rightarrow E$ relations in Lesson 2, and 96 relations involving combinations of D and B (or C) stimuli as samples and either E or F stimuli as comparisons in Lesson 3. All told, teaching 40 relations resulted in up to 144 emergent relations (the yield was slightly lower for students who showed the unusual error patterns shown in Figure 4). Thus, the lessons were capable of promoting as many as 4.7 times as many relations as were directly taught, thereby illustrating the instructional efficiency that is a hallmark of EBI (Critchfield & Fienup, 2008; Stromer, Mackay, & Stoddard, 1992).

The present study joins those of Ninness et al. (2005, 2006, 2009) and Fields et al. (2009) as an extension of EBI to advanced academic

subjects as learned by sophisticated students. These studies are important because, as learners progress from public education to postsecondary and professional education, the amount of time allotted to formal instruction tends to decrease. For example, elementary students spend 6 to 8 hr per day in class; by contrast, many college students spend perhaps a few hours per week. Thus, even in advanced academic programs, it is important to get the most out of limited instructional time (e.g., Chew, 2008).

From a research perspective, the present study provides a more straightforward demonstration of the effects of EBI with advanced learners than previous studies by Ninness et al. (2005, 2006), because those studies also incorporated instructor-generated explanations of the rules on how stimuli were related. The present study offers the advantage of showing how much academic gain students derived from relatively little practice with the component relations, even without instructor-generated explanations of underlying rules regarding stimulus relations.

Contextual Control

Although many previous studies have illustrated the academic promise of EBI, to our knowledge the present study represents the first attempt to employ contextually controlled equivalence classes in a program of instruction. In Lesson 2, students learned to make hypothesis decisions based solely on whether the direction of an effect corresponded to the scientific hypothesis prediction. Lesson 3 restructured and enhanced this repertoire. In cases in which effects correspond to the scientific hypothesis prediction (but not in cases in which effects run counter to that prediction), hypothesis decisions require consideration of statistical information. Thus, statistical information serves a contextual function, signaling how effects guide hypothesis decisions. For all students, contextual control was promoted during Lesson 3 when the A stimuli (*low p value* and *high p*

value) came to modulate how combinations of effects and scientific hypotheses guided hypothesis decisions.

It should be noted that in most previous studies of contextual control, the contextual stimuli served *only* a contextual function (e.g., Bush et al., 1986; Gatch & Osborne, 1989; Hayes, Kohlenberg, & Hayes, 1991), but in the present study, the contextual stimulus (A) of Lesson 3 also was a member of an independent equivalence class (developed during Lesson 1). Consequently, for 9 of 10 students its associates (B and C) also came to serve as contextual stimuli without any specific training to promote this function. Transformation of function (also called transfer of function), in which a function served by one class member is acquired spontaneously by other class members, appears to be a hallmark of emergent stimulus classes (Dymond, 2000). At least one laboratory study has shown that contextual cuing functions can propagate through equivalence classes in this way (Hayes et al., 1991), an effect that the present lessons apparently replicated.

Although some students experienced no difficulty in rendering hypothesis decisions in the presence of statistical information, others did experience difficulty. As Figure 4 illustrates, for trials on which the D stimulus described a mismatch between the scientific hypothesis and the direction of an effect, the B stimulus (*statistically significant*) and C stimulus ($p > .05$) continued to be treated as a member of the original A-B-C classes when presented as part of the sample stimulus. This was not always true, however, when the B stimulus (five students) and C stimulus (four students) served as a comparison.

To our knowledge such highly selective fracturing of equivalence classes has no precedent in either the basic or applied stimulus equivalence literature on contextual control, not the least because most studies have lacked a standardized means of measuring collateral change of function in contextual stimuli.

Skinner (1968) argued that the process of programming instruction provides a means of detecting skill gaps that might require extra attention, because instructional programming demands a precise task analysis of the skills that students must master. The analysis simultaneously identifies what must be taught and what should be assessed as evidence of learning. In the present study, the task analysis was guided partly by the stimulus equivalence framework, in which concepts are understood in terms of component relations (including emergent ones), and partly by studies showing that equivalence classes can sometimes be disrupted when the function of one member of a class is altered (Pilgrim & Galizio, 1995), as was the case during Lesson 3 contextual training. This led to the creation of the maintenance tests that followed Lesson 3 and that revealed interesting error patterns for some students. In evaluating student success, had we relied on global percentage correct scores for the tests (as often is done in academic settings) we might have overlooked isolated special needs that could be addressed easily through follow-up instruction on an individual basis.

Future Directions

A general shortcoming of the existing scientific literature on EBI is that most studies were conducted under highly controlled conditions. Although research shows that this type of instruction *can* generate academically relevant skills, for the most part it remains to be seen whether such gains *will* occur under a given set of everyday instructional circumstances (e.g., Chorpita, 2003). In addition, whether EBI really is more efficient than other techniques has rarely been tested (for an example, see Taylor & O'Reilly, 2000).

Future research could address why students evidenced isolated pockets of misunderstanding (Students 3, 5, 6, and 10 on the A-B-C maintenance tests that followed Lesson 3). We have not identified any factors that might have placed these particular students at risk for less-

than-ideal outcomes. Like other students, they scored low on pretests and, following training, high on posttests. During all training phases, they progressed about as quickly as other students. In short, these students appeared to be unremarkable in most ways, yet their special difficulties may highlight the inexact fit of standardized instruction to the needs of a heterogeneous student population. However, these problems may have been anticipated based on what is known about stimulus control. Recall that the relations taught in Lesson 2 were not conditional (students were taught to evaluate hypotheses based only on the correspondence between scientific hypothesis prediction and direction of effects, without incorporating inferential statistics information). Later training (Lesson 3) modified some of what had been learned (reversed relations) to create the required conditional reasoning. Two principles of stimulus control may be relevant here. Based on the literature of errorless discrimination learning (e.g., Terrace, 1963), better outcomes are predicted for procedures in which key discriminations are introduced as early as possible during instruction. Based on a few studies in which already-established equivalence classes were altered (e.g., Pilgrim & Galizio, 1995), perhaps we might have expected the A-B-C class of Lesson 1 to be compromised by Lesson 3 contextual training in which the functions of selected stimuli were redefined.

In developing the present lessons, we considered establishing contextually controlled relations from the start. That is, rather than building incomplete repertoires during Lesson 2 and then correcting these repertoires in Lessons 3, the two lessons could be combined into one omnibus training session. A potential drawback of this approach is to require a fairly extensive training regimen prior to testing, and the present training sequence was adopted after pilot work suggested that its relatively brief training phases had favorable effects on student motivation and attention. We did not, however,

formally compare the two approaches, so it remains possible that better outcomes would be obtained using a different training structure. Such a comparison is the focus of research now in progress.

The preceding discussion highlights the complexity of issues that must be considered in judging whether an intervention like the present one is ready for application in natural settings. On the one hand, traditional applied behavioral interventions typically are thought of as serving individuals, in which case the progress of every individual is equally important (e.g., Johnston & Pennypacker, 1993). Thus, error patterns like those of the aforementioned four students should concern the instructional designer because they represent potential weaknesses in instructional stimulus control. From this perspective, the present lessons may require additional development before further application is warranted. On the other hand, interventions designed for simultaneous use with many individuals (as might be encountered in group academic settings) typically do not attempt to “save” everyone but rather to use available resources to create good outcomes for as many individuals as possible. In such interventions, it is a good day when most students learn most of what was taught—all the more if not everything learned had to be taught directly. From this perspective, it will be interesting to see what outcomes could be achieved by employing the present lessons with large numbers of students under natural instructional contingencies.

REFERENCES

- Belanich, J., & Fields, L. (2003). Generalized equivalence classes as response transfer networks. *The Psychological Record*, 53, 373–413.
- Bush, K. M., Sidman, M., & de Rose, T. (1989). Contextual control of emergent equivalence relations. *Journal of the Experimental Analysis of Behavior*, 51, 29–45.
- Chew, S. L. (2008). Study more! Study harder! Students' and teachers' faulty beliefs about how people learn. In S. A. Meyers & J. R. Stowell (Eds.), *Essays from excellence in teaching*. (Vol. 7, pp. 22–25). Retrieved from: <http://teachpsych.org/resources/e-books/eit2007/eit2007.php>
- Chorpita, B. F. (2003). The frontier of evidence-based practice. In A. E. Kazdin & J. R. Weisz (Eds.), *Evidence-based psychotherapies for children and adolescents* (pp. 42–59). New York: Guilford.
- Connell, J. E., & Witt, J. C. (2004). Applications of computer-based instruction: Using specialized software to aid letter-name and letter-sound recognition. *Journal of Applied Behavior Analysis*, 37, 67–71.
- Critchfield, T. S., & Fienup, D. M. (2008). Stimulus equivalence. In S. F. Davis & W. F. Buskist (Eds.), *21st century psychology* (pp. 360–372). Thousand Oaks, CA: Sage.
- de Rose, J. C., de Souza, D. G., & Hanna, E. S. (1996). Teaching reading and spelling: Exclusion and stimulus equivalence. *Journal of Applied Behavior Analysis*, 29, 451–469.
- Dixon, M. R., & MacLin, O. H. (2003). *Visual basic for behavioral psychologists*. Reno, NV: Context Press.
- Dougher, M., Perkins, D. R., Greenway, D., Koons, A., & Chiasson, C. (2002). Contextual control of equivalence-based transformation of functions. *Journal of the Experimental Analysis of Behavior*, 78, 63–93.
- Dymond, S. (2000). Understanding complex behavior: The transformation of stimulus functions. *The Behavior Analyst*, 23, 239–254.
- Fields, L., Travis, R., Roy, D., Yadlovker, E., deAguiar-Rocha, L., & Sturmey, P. (2009). Equivalence class formation: A method for teaching statistical interactions. *Journal of Applied Behavior Analysis*, 42, 575–593.
- Fienup, D. M., Critchfield, T. S., & Covey, D. P. (2009). Building contextually-controlled equivalence classes to teach about inferential statistics: A preliminary demonstration. *Experimental Analysis of Human Behavior Bulletin*, 27, 1–10.
- Fienup, D. M., & Dixon, M. R. (2006). Acquisition and maintenance of visual-visual and visual-olfactory equivalence classes. *European Journal of Behavior Analysis*, 7, 87–98.
- Gatch, M. B., & Osborne, J. G. (1989). Transfer of contextual stimulus function via equivalence class development. *Journal of the Experimental Analysis of Behavior*, 51, 369–378.
- Griffiee, K., & Dougher, M. J. (2002). Contextual control of stimulus generalization and stimulus equivalence in hierarchical categorization. *Journal of the Experimental Analysis of Behavior*, 78, 433–447.
- Hayes, S. C., Kohlenberg, B. S., & Hayes, L. J. (1991). The transfer of specific and general consequential functions through simple and conditional equivalence relations. *Journal of the Experimental Analysis of Behavior*, 56, 119–137.
- Huck, S. W. (2000). *Reading statistics and research* (3rd ed.). New York: Longman.

Johnston, J. M., & Pennypacker, H. S. (1993). *Strategies and tactics of behavioral research* (2nd ed.). Hillsdale, NJ: Erlbaum.

Keenan, M., McGlinchey, A., Fairhurst, C., & Dillenburger, K. (2000). Accuracy of disclosure and contextual control in child abuse: Developing procedures within the stimulus equivalence paradigm. *Behavior and Social Issues, 10*, 1–17.

Knowles, L. (1974). Helping students learn basic inferential statistics. *College Student Journal, 8*, 7–11.

Kohlenberg, B. S., Hayes, S. C., & Hayes, L. J. (1991). The transfer of contextual control over equivalence classes through equivalence classes: A possible model of social stereotyping. *Journal of the Experimental Analysis of Behavior, 56*, 505–518.

Kranzler, J. H. (2007). *Statistics for the terrified* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Lane, S. D., Clow, J. K., Innis, A., & Critchfield, T. S. (1998). Generalization of cross-modal stimulus equivalence classes: Operant processes as components in human category formation. *Journal of the Experimental Analysis of Behavior, 70*, 267–279.

Lane, S. D., & Critchfield, T. S. (1998). Classification of vowels and consonants by individuals with moderate mental retardation: Development of arbitrary relations via match-to-sample training with compounds. *Journal of Applied Behavior Analysis, 31*, 21–41.

Lynch, D. C., & Cuvo, A. J. (1995). Stimulus equivalence instruction of fraction-decimal relations. *Journal of Applied Behavior Analysis, 28*, 115–126.

Mattaini, M. A. (1999). Letter to the editor: One more clinical implication of stimulus equivalence research. *Behavior Therapy, 30*, 341–343.

Ninness, C., Barnes-Holmes, D., Rumph, R., McCullen, G., Ford, A. M., Payne, R., et al. (2006). Transformation of mathematical and stimulus functions. *Journal of Applied Behavior Analysis, 39*, 299–321.

Ninness, C., Dixon, M., Barnes-Holmes, D., Rehfeldt, R. A., Rumph, R., McCullen, G., et al. (2009). Constructing and deriving reciprocal trigonometric relations: A functional analytic approach. *Journal of Applied Behavior Analysis, 42*, 191–208.

Ninness, C., Rumph, R., McCullen, G., Harrison, C., Ford, A. M., & Ninness, S. K. (2005). Functional analytic approach to computer-interactive mathematics. *Journal of Applied Behavior Analysis, 38*, 1–22.

Perone, M. (2002). Behavior analysis and translational research. *ABA Newsletter, 25*, (2), 3–4.

Pilgrim, C., & Galizio, M. (1995). Reversal of baseline relations and stimulus equivalence: I. Adults. *Journal of the Experimental Analysis of Behavior, 63*, 225–238.

Rehfeldt, R. A. (2003). Establishing contextual control over generalized equivalence relations. *The Psychological Record, 53*, 415–428.

Sidman, M. (1994). *Equivalence relations and behavior: A research story*. Boston: Authors Cooperative.

Sidman, M., & Cresson, O. (1973). Reading and crossmodal transfer of stimulus equivalence in severe retardation. *American Journal of Mental Deficiency, 77*, 515–523.

Sidman, M., Kirk, B., & Willson-Morris, M. (1985). Six-member stimulus classes generated by conditional-discrimination procedures. *Journal of the Experimental Analysis of Behavior, 43*, 21–42.

Skinner, B. F. (1968). *The technology of teaching*. New York: Appleton-Century-Crofts.

Steele, D., & Hayes, S. C. (1991). Stimulus equivalence and arbitrarily applicable relational responding. *Journal of the Experimental Analysis of Behavior, 56*, 519–555.

Stromer, R., Mackay, H., & Stoddard, L. (1992). Classroom applications of stimulus equivalence technology. *Journal of Behavioral Education, 2*, 225–256.

Taylor, I., & O'Reilly, M. F. (2000). Generalization of supermarket shopping skills for individuals with mild intellectual disabilities using stimulus equivalence training. *The Psychological Record, 50*, 49–62.

Terrace, H. S. (1963). Discrimination learning with and without “errors.” *Journal of the Experimental Analysis of Behavior, 6*, 1–27.

Received June 17, 2008
Final acceptance January 19, 2010
Action Editor, Chris Ninness

APPENDIX A
Lesson 1: Trained A-B-C Relations

| Phase | Sample stimulus | | Comparison stimuli | | | |
|-------|-----------------|---------------------|--------------------|-------------------------------|----------|-------------------------------|
| | Notation | Stimulus | Notation | Correct choice | Notation | Incorrect choice |
| 1 | A1 | Low <i>p</i> value | B1 | Statistically significant | B2 | Not statistically significant |
| 1 | A2 | High <i>p</i> value | B2 | Not statistically significant | B1 | Statistically significant |
| 2 | C1 | <i>p</i> ≤ .05 | A1 | Low <i>p</i> value | A2 | High <i>p</i> value |
| 2 | C2 | <i>p</i> > .05 | A2 | High <i>p</i> value | A1 | Low <i>p</i> value |

Note. Presented are the relations that were explicitly trained during Lesson 1 (see text for details). Each row summarizes a trial type and the phase during which that trial was presented. For each trial the sample stimulus and corresponding correct and incorrect comparison stimuli are presented.

APPENDIX B
Lesson 2: Trained D-E-F Relations

| Phase | Sample stimulus | | Comparison stimuli | | | |
|-------|-----------------|--|--------------------|---|----------|---|
| | Notation | Stimulus | Notation | Correct choice | Notation | Incorrect choice |
| 1 | ↑ D1 | Scientific hypothesis: the IV will increase the DV Results: the IV increased | E1 | Consistent with the scientific hypothesis | E2 | Not consistent with the scientific hypothesis |
| 1 | ↑ D2 | Scientific hypothesis: the IV will increase the DV Results: the IV did not increase | E2 | Not consistent with the scientific hypothesis | E1 | Consistent with the scientific hypothesis |
| 2 | ↑ D1 | Scientific hypothesis: the IV will increase the DV Results: the IV increased | F1 | Reject the null hypothesis | F2 | Fail to reject the null hypothesis |
| 2 | ↑ D2 | Scientific hypothesis: the IV will increase the DV Results: the IV did not increase | F2 | Fail to reject the null hypothesis | F1 | Reject the null hypothesis |
| 3 | ↓ D1 | Scientific hypothesis: the IV will decrease the DV Results: the IV decreased | E1 | Consistent with the scientific hypothesis | E2 | Not consistent with the scientific hypothesis |
| 3 | ↓ D2 | Scientific hypothesis: the IV will decrease the DV Results: the IV did not decrease | E2 | Not consistent with the scientific hypothesis | E1 | Consistent with the scientific hypothesis |
| 4 | ↓ D1 | Scientific hypothesis: the IV will decrease the DV Results: the IV decreased | F1 | Reject the null hypothesis | F2 | Fail to reject the null hypothesis |
| 4 | ↓ D2 | Scientific hypothesis: the IV will decrease the DV Results: the IV did not decrease | F2 | Fail to reject the null hypothesis | F1 | Reject the null hypothesis |
| 5 | ↕ D1 | Scientific hypothesis: the IV will change the DV Results: the IV changed | E1 | Consistent with the scientific hypothesis | E2 | Not consistent with the scientific hypothesis |
| 5 | ↕ D2 | Scientific hypothesis: the IV will change the DV Results: the IV did not change | E2 | Not consistent with the scientific hypothesis | E1 | Consistent with the scientific hypothesis |
| 6 | ↕ D1 | Scientific hypothesis: the IV will change the DV Results: the IV changed | F1 | Reject the null hypothesis | F2 | Fail to reject the null hypothesis |
| 6 | ↕ D2 | Scientific hypothesis: the IV will change the DV Results: the IV did not change | F2 | Fail to reject the null hypothesis | F1 | Reject the null hypothesis |

Note. Presented are the relations that were explicitly trained during Lesson 2 (see text for details). Each row summarizes a trial type and the phase during which that trial was presented. For each trial the sample stimulus and corresponding correct and incorrect comparison stimuli are presented.

APPENDIX C
Lesson 3: Trained Contextual Relations

| Phase | Sample stimulus | | Comparison stimuli | | | |
|-------|-----------------|--|--------------------|---|----------|---|
| | Notation | Stimulus | Notation | Correct choice | Notation | Incorrect choice |
| 1 | ↑ D1/A1 | Scientific hypothesis: the IV will increase the DV Results: the IV increased ----- Low <i>p</i> value | E1 | Consistent with the scientific hypothesis | E2 | Not consistent with the scientific hypothesis |
| 1 | ↑ D2/A1 | Scientific hypothesis: the IV will increase the DV Results: the IV did not increase ----- Low <i>p</i> value | E2 | Not consistent with the scientific hypothesis | E1 | Consistent with the scientific hypothesis |
| 2 | ↑ D1/A2 | Scientific hypothesis: the IV will increase the DV Results: the IV increased ----- High <i>p</i> value | E2 | Not consistent with the scientific hypothesis | E1 | Consistent with the scientific hypothesis |
| 2 | ↑ D2/A2 | Scientific hypothesis: the IV will increase the DV Results: the IV did not increase ----- High <i>p</i> value | E2 | Not consistent with the scientific hypothesis | E1 | Consistent with the scientific hypothesis |
| 3 | ↑ D1/A1 | Scientific hypothesis: the IV will increase the DV Results: the IV increased ----- Low <i>p</i> value | F1 | Reject the null hypothesis | F2 | Fail to reject the null hypothesis |
| 3 | ↑ D2/A1 | Scientific hypothesis: the IV will increase the DV Results: the IV did not increase ----- Low <i>p</i> value | F2 | Fail to reject the null hypothesis | F1 | Reject the null hypothesis |
| 4 | ↑ D1/A2 | Scientific hypothesis: the IV will increase the DV Results: the IV increased ----- High <i>p</i> value | F2 | Fail to reject the null hypothesis | F1 | Reject the null hypothesis |
| 4 | ↑ D2/A2 | Scientific hypothesis: the IV will increase the DV Results: the IV did not increase ----- High <i>p</i> value | F2 | Fail to reject the null hypothesis | F1 | Reject the null hypothesis |
| 5 | ↓ D1/A1 | Scientific hypothesis: the IV will decrease the DV Results: the IV decreased ----- Low <i>p</i> value | E1 | Consistent with the scientific hypothesis | E2 | Not consistent with the scientific hypothesis |

APPENDIX C

(Continued)

| Phase | Sample stimulus | | Comparison stimuli | | | |
|-------|-----------------|---|--------------------|--|----------|--|
| | Notation | Stimulus | Notation | Correct choice | Notation | Incorrect choice |
| 5 | ↓ D2/A1 | Scientific hypothesis: the IV will decrease the DV Results: the IV did not decrease ----- Low p value | E2 | Not consistent with the scientific hypothesis | E1 | Consistent with the scientific hypothesis |
| 6 | ↓ D1/A2 | Scientific hypothesis: the IV will decrease the DV Results: the IV decreased ----- High p value | E2 | Not consistent with the scientific hypothesis | E1 | Consistent with the scientific hypothesis |
| 6 | ↓ D2/A2 | Scientific hypothesis: the IV will decrease the DV Results: the IV did not decrease ----- High p value | E2 | Not consistent with the scientific hypothesis | E1 | Consistent with the scientific hypothesis |
| 7 | ↓ D1/A1 | Scientific hypothesis: the IV will decrease the DV Results: the IV decreased ----- Low p value | F1 | Reject the null hypothesis | F2 | Fail to reject the null hypothesis |
| 7 | ↓ D2/A1 | Scientific hypothesis: the IV will decrease the DV Results: the IV did not decrease ----- Low p value | F2 | Fail to reject the null hypothesis | F1 | Reject the null hypothesis |
| 8 | ↓ D1/A2 | Scientific hypothesis: the IV will decrease the DV Results: the IV decreased ----- High p value | F2 | Fail to reject the null hypothesis | F1 | Reject the null hypothesis |
| 8 | ↓ D2/A2 | Scientific hypothesis: the IV will decrease the DV Results: the IV did not decrease ----- High p value | F2 | Fail to reject the null hypothesis | F1 | Reject the null hypothesis |
| 9 | ↑ D1/A1 | Scientific hypothesis: the IV will change the DV Results: the IV changed ----- Low p value | E1 | Consistent with the scientific hypothesis | E2 | Not consistent with the scientific hypothesis |
| 9 | ↑ D2/A1 | Scientific hypothesis: the IV will change the DV Results: the IV did not change ----- Low p value | E2 | Not consistent with the scientific hypothesis | E1 | Consistent with the scientific hypothesis |

APPENDIX C
(Continued)

| Phase | Sample stimulus | | Comparison stimuli | | | |
|-------|-----------------|--|--------------------|---|----------|---|
| | Notation | Stimulus | Notation | Correct choice | Notation | Incorrect choice |
| 10 | ↕ D1/A2 | Scientific hypothesis: the IV will change the DV Results: the IV changed ----- High <i>p</i> value | E2 | Not consistent with the scientific hypothesis | E1 | Consistent with the scientific hypothesis |
| 10 | ↕ D2/A2 | Scientific hypothesis: the IV will change the DV Results: the IV did not change ----- High <i>p</i> value | E2 | Not consistent with the scientific hypothesis | E1 | Consistent with the scientific hypothesis |
| 11 | ↕ D1/A1 | Scientific hypothesis: the IV will change the DV Results: the IV changed ----- Low <i>p</i> value | F1 | Reject the null hypothesis | F2 | Fail to reject the null hypothesis |
| 11 | ↕ D2/A1 | Scientific hypothesis: the IV will change the DV Results: the IV did not change ----- Low <i>p</i> value | F2 | Fail to reject the null hypothesis | F1 | Reject the null hypothesis |
| 12 | ↕ D1/A2 | Scientific hypothesis: the IV will change the DV Results: the IV changed ----- High <i>p</i> value | F2 | Fail to reject the null hypothesis | F1 | Reject the null hypothesis |
| 12 | ↕ D2/A2 | Scientific hypothesis: the IV will change the DV Results: the IV did not change ----- High <i>p</i> value | F2 | Fail to reject the null hypothesis | F1 | Reject the null hypothesis |

Note. Presented are the relations that were explicitly trained during Lesson 3 (see text for details). Each row summarizes a trial type and the phase during which that trial was presented. For each trial the sample stimulus and corresponding correct and incorrect comparison stimuli are presented.